Cours de Statistiques

pour la description de données

Romain Raveaux[1]

[1]Laboratoire L3I – Université de La Rochelle
romain.raveaux01 at univ-lr.fr

Octobre 24-11, 2008

Content

Type de variable

Numérique

- Soit l'étude de la variable X, une série de valeurs définies dans $\Re$.
- Exemple: Age, poids,...

Nominale

- Ne prend qu'un nombre limité de valeurs.
- Et que ces valeurs n'ont entre elles aucune relation apparente.
- Exemple : Le statut marital, qui pourrait prendre les valeurs "Célibataire", "Marié", "Veuf", "Divorcé", "Union libre".

Ordinale

- Ne prend qu'un nombre limité de valeurs.
- Et que ces valeurs n'ont entre elles aucune relation apparente.
- Les grades dans l'armée: "lieutenant", "capitaine", "commandant" etc...
- Par nature, les rangs sont des variables ordinales.

Il existe d'autres types de variable : Binaire, Normale,...

Variable et Espace d'étude

Une série à valeurs individuelles

- Soit l'étude de la variable X, une série de valeurs définies dans $\Re$.

Statistiques multi-dimensionnelles

- Soit l'étude d'un ensemble fini de variables (Ω), Ω est l'univers des statistiques.
- Avec $card(\Omega) = M$
- $\Omega = X_1, X_2, ..., X_m$
- $\forall X_i \in \Omega$, X_i est une série à valeurs individuelles.

Estimateurs

Soit l'étude de la variable X, une série de valeurs définies dans $\Re^+$:

Moyenne d'une série à valeurs individuelles

- $\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$

Variance d'une série à valeurs individuelles

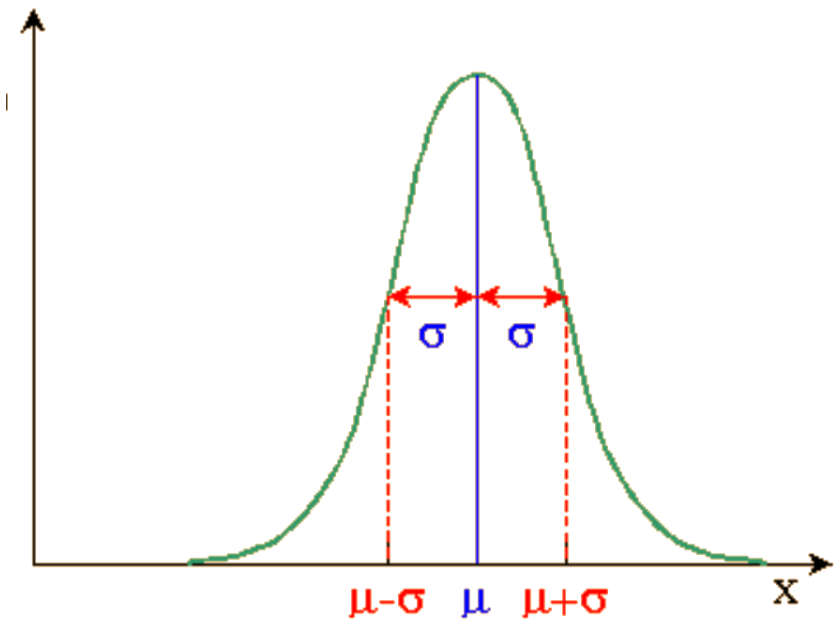
- $\theta^2 = V(X) = \sum_{i=1}^{N} (x_i - \overline{X})^2$
- L'ecart type ce déduit de la variance : $\theta = \sqrt{V(X)}$

Représentation de ces estimateurs

Soit l'étude de la variable X suivant une loi normale ($\aleph(\mu, \theta^2)$), de moyenne μ et de variance θ.

Densité de probabilité d'une loi gaussienne

- $f(x) = \frac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\theta})^2}$

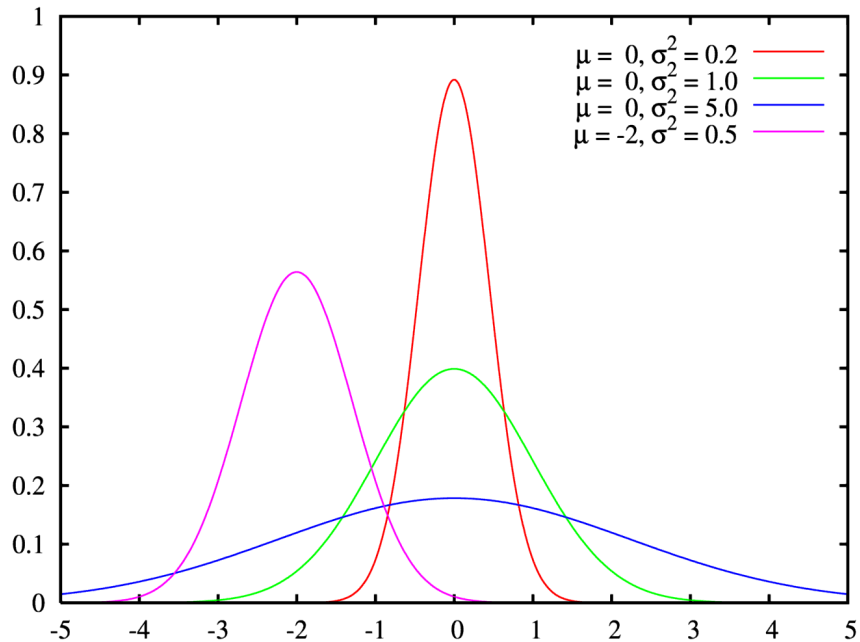


Représentation de ces estimateurs

Soit l'étude de la variable X suivant une loi normale $(\aleph(\mu, \theta^2))$, de moyenne μ et de variance θ.

Densité de probabilité d'une loi gaussienne

- $f(x) = \frac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\theta})^2}$



Explication intuitive de ces estimateurs

Plus la variance d'un échantillon est grande et plus les données sont éparses. Cela peut dénoter une érreur dans le phénoméne mesuré.

Relations entre deux séries de données observées

Exemples :

- Mesurer le poids ou la longueur d'un organe (variable dépendante) à différentes dates successives choisies arbitrairement (variable indépendante).
- Mesurer le rendement d'une culture (variable dépendante) en fonction de différentes doses d'engrais (variable indépendante).
- Mesurer la capacité à résoudre un problème ou à réaliser une tâche (variable dépendante) en fonction de différentes doses d'un médicament (variable indépendante).

Covariance de deux échantillons

Soit l'étude de deux variables X et Y, deux séries de valeurs définies dans $\Re$:

Covariance

- $\theta_{xy} = cov(X, Y) = \sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})$
- La fonction covariance retourne des valeurs comprises dans $[-\infty, +\infty]$
- X et Y indépendant $\Longrightarrow cov(X, Y) = 0$

Covariance de deux échantillons

Covariance

- $\theta_{xy} = cov(X, Y) = \sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})$
- Intuitivement, la covariance est une mesure de la variation simultanée de deux variables aléatoires.
- C'est-à-dire que la covariance devient plus positive pour chaque couple de valeurs qui diffèrent de leur moyenne dans le même sens, et plus négative pour chaque couple de valeurs qui diffèrent de leur moyenne dans le sens opposé.

Corrélation de deux variables aléatoires

Soit l'étude de deux variables X et Y, deux séries de valeurs définies dans $\Re$:

Corrélation de Bravais-Pearson

- $cor(X,Y) = \frac{\theta_{xy}}{\theta_x.\theta_y} = \frac{cov(X,Y)}{\sqrt{cov(X)}.\sqrt{cov(Y)}}$
- Le coéfficient de corrélation est compris entre $[-1,1]$
- $cor(X,Y) = 0 \Longrightarrow$, X et Y sont indépendant linéairement.
- $cor(X,Y) = 1$, une relation affine existe entre X et Y. L'une des variables est fonction affine croissante de l'autre variable.
- $cor(X,Y) = -1$, une relation affine existe entre X et Y. L'une des variables est fonction affine décroissante de l'autre variable.

Corrélation de Kendall

Soit deux variables ordinales X et Y. La corrélation de rangs rend compte d'une relation non-linéaire entre ces deux variables.
τ s'exprime de la façon suivante :

$$\tau = \frac{S}{D}$$

Où,

$$S = \sum_{i<j} (sign(x[i] - y[i]).sign(y[i] - x[i])) \quad (1)$$

et,

$$D = \frac{k(k-1)}{2} \quad (2)$$

Tableau de contingence

Prenons le temps de faire un petit sondage anonyme au sein de la classe :

	Université	IUT/BTS	Autre	Total
L	?	?	?	..
ES	?	?	?	..
S	?	?	?	..
ST	?	?	?	..
Total	..	..	..	N

Pour être appelé tableau de contingence, il faut pour cela que les nombres dans les cellules soient le résultat d'un décompte, de façon à ce que additionner les contenus des cellules d'une ligne ou d'une colonne ait un sens.

Question

Quelle matrice aurait-on si la répartition dans les filières post-Bac ne dépendait pas du type de Bac ?

Notion d'indépendance

	Université	IUT/BTS	Autre	Total
L	13	2	5	20
ES	20	2	8	30
S	10	5	5	20
ST	7	1	22	30
Total	50	10	40	100

Quelle matrice aurait-on si la répartition dans les filières post-Bac ne dépendait pas du type de Bac ?

Notion d'indépendance

Quelle matrice aurait-on si la répartition dans les filières post-Bac ne dépendait pas du type de Bac ?

	Université	IUT/BTS	Autre	Total
L	13	2	5	20
ES	20	2	8	30
S	10	5	5	20
ST	7	1	22	30
Total	50	10	40	100

On reconstitue la matrice à partir de ses marges :

	Université	IUT/BTS	Autre	Total
L	50*20/100	10*20/100	8	20
ES	50*30/100	13	12	30
S	10	2	8	20
ST	15	13	12	30
Total	50	10	40	100

Tableau de contingence

Une entreprise vend 5 produits dans 4 régions. A la fin de chaque exercice, ses ventes, exprimées par exemple en milliers d'unités, peuvent se résumer dans un tableau comme celui-ci :

	P1	P2	P3	P4	P5
R1	28	14	45	33	12
R2	36	21	25	64	23
R3	21	64	38	11	7
R4	79	42	67	9	41

Pour être appelé tableau de contingence, il faut pour cela que les nombres dans les cellules soient le résultat d'un décompte, de façon à ce que additionner les contenus des cellules d'une ligne ou d'une colonne ait un sens.

Tableau de contingence

Une entreprise vend 5 produits dans 4 régions. A la fin de chaque exercice, ses ventes, exprimées par exemple en milliers d'unités, peuvent se résumer dans un tableau comme celui-ci :

	P1	P2	P3	P4	P5
R1	28	14	45	33	12
R2	36	21	25	64	23
R3	21	64	38	11	7
R4	79	42	67	9	41

L'analyse des interactions entre deux variables nominales. Dans l'exemple ci-dessus, nous avons deux telles variables :

- Produit (5 modialités)
- Région (4 modialités)

Notations

- n_{ij} : effectif de la cellule (i,j),
- $n_{i\bullet}$: effectif total de la ligne i,
- $n_{\bullet j}$: effectif total de la colonne j,
- $n_{\bullet\bullet}$: effectif total.

Tableau des fréquences

Les fréquences sont calculées par :

$$f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}} = \frac{Effectif\ de\ lacellule(i,j)}{Effectif\ total}$$

Tableau des fréquences lignes

Les fréquences lignes (ou coordonnées des profils lignes) sont calculées par :

$$fl_{ij} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}} = \frac{Effectif\ de\ la cellule(i,j)}{Effectif\ total\ de\ la\ ligne\ i}$$

Les coordonnées du profil ligne moyen sont calculées par :

$$f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$$

Tableau des fréquences colonnes

Les fréquences colonnes sont calculées par :

$$fc_{ij} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}} = \frac{Effectif\ de\ lacellule(i,j)}{Effectif\ total\ de\ la\ colonnes\ j}$$

Les coordonnées du profil colonne moyen sont calculées par :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}}$$

Distances entre profils

Distance entre deux lignes. Métrique du χ^2

$$d_{\chi^2}(L_1, L_2) = \sum_j \frac{(fl_{1j} - fl_{2j})^2}{f_{\bullet j}}$$

Distance entre deux lignes. Métrique Euclidienne

$$d_{\chi^2}(L_1, L_2) = \sum_j (fl_{1j} - fl_{2j})^2$$

Taux de liaison

Représente les écarts entre les valeurs théoriques en cas d'indépendance et les valeurs observées

$$t_{ij} = \frac{f_{ij} - f_{i\bullet}.f_{\bullet j}}{f_{i\bullet}.f_{\bullet j}}$$

Distance entre deux variables nominales

Représente les écarts entre les valeurs théoriques en cas d'indépendance et les valeurs observées

$$\chi^2 = \sum_{ij} \frac{(f_{ij} - f_{i\bullet}.f_{\bullet j})^2}{f_{i\bullet}.f_{\bullet j}}$$

Variables indépendantes

	P1	P2	P3	P4	P5
R1	28	14	45	33	12
R2	36	21	25	64	23
R3	21	64	38	11	7
R4	79	42	67	9	41

Question : Les deux variables "Région" et "Produit" sont indépendantes ? Explication : Ceci voudrait dire que deux régions quelconques vendent tous les produits exactement dans les mêmes proportions ou/et que deux produits quelconques sont vendus exactement dans les mêmes proportions dans toutes les régions.

Variables indépendantes

	P1	P2	P3	P4	P5
R1	28	14	45	33	12
R2	36	21	25	64	23
R3	21	64	38	11	7
R4	79	42	67	9	41

Réponse intuitive : Par exemple, nous voyons que la Région R1 vend deux fois plus de produits P1 (28) que de produits P2 (14). Si "Région" et "Produit" étaient indépendants, nous nous attendrions à ce qu'il en soit de même pour la région R2.Mais nous constatons que R2 a vendu 36 P1 et 21 P2, et non les $36/2 = 18$ attendus.

Variables indépendantes

	P1	P2	P3	P4	P5
R1	28	14	45	33	12
R2	36	21	25	64	23
R3	21	64	38	11	7
R4	79	42	67	9	41

Forcé de constater qu'une approche visiuelle est inapplicable à de grands tableaux et de plus l'écart entre 21 (observé) 18 (theorique) est faible. Cette petite différence ne serait-elle pas simplement imputable aux fluctuations naturelles de la vie normale de l'entreprise ? Plus généralement, comment être sûr que les deux variables ne sont pas indépendantes, puisque nous ne pouvons pas complètement faire confiance aux chiffres ?

Test du χ^2

Question : Les deux variables "Région" et "Produit" sont indépendantes ? Ce genre de question appelle clairement un test. Dans le vocabulaire standard des test, l'hypothèse nulle H0 est : "Les deux variables sont indépendantes". Le "test du Chi-deux d'indépendance" va construire une quantité :

- égale à "0" quand les nombres du tableau correspondent parfaitement à ce que l'on attendrait en cas d'indépendence des deux variables,
- positive autrement,
- et qui devient de plus en plus grande au fur et à mesure que la distribution observée s'écarte de la distribution idéale en cas d'indépendance.

Test du χ^2

Question : Les deux variables "Région" et "Produit" sont indépendantes ?
Le test calculera alors la probabilité pour que cette quantité soit encore plus grande que celle effectivement observée si les deux variables sont effectivement indépendantes. Une faible valeur de cette probabilité (p. ex. 0,005) plaide en faveur de l'existence d'un lien entre ces variables, et donc est un quasi-démenti à l'hypothèse d'indépendance.

Contribution au χ^2

Les deux caractères sont x et y, la taille de l'échantillon est n. Les modalités ou classes de x seront notées $c_1, \ldots, c_r$, celles de y sont notées $d_1, \ldots, d_s$. On note :

- n_{hk} l'effectif conjoint de c_h et d_k : c'est le nombre d'individus pour lesquels x prend la valeur c_h et y la valeur d_k,
- $n_{h\bullet} = \sum_{k=1}^{s} n_{hk}$ l'effectif marginal de c_h : c'est le nombre d'individus pour lesquels x prend la valeur c_h,
- $n_{\bullet k} = \sum_{h=1}^{r} n_{hk}$ l'effectif marginal de d_k : c'est le nombre d'individus pour lesquels y prend la valeur d_k.

Contribution au χ^2

On représente ces valeurs dans un tableau à double entrée, dit tableau de contingence.

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

En divisant les lignes et les colonnes par leurs sommes, on obtient sur chacune des distributions empiriques constituées de fréquences conditionnelles.

Contribution au χ^2

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

Pour $h = 1, \ldots, r$ et $k = 1, \ldots, s$, on les notera :

$$f_{k|h} = \frac{n_{hk}}{n_{h\bullet}} \text{ et } \quad f_{h|k} = \frac{n_{hk}}{n_{\bullet k}}$$

Ces distributions empiriques conditionnelles s'appellent les profils-lignes et profils-colonnes.

Contribution au χ^2

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

- L'enjeu principal est d'étudier la dépendance des deux caractères.
- Deux caractères sont indépendants si la valeur de l'un n'influe pas sur les distributions des valeurs de l'autre.

Si c'est le cas, les profils-lignes seront tous peu différents de la distribution empirique de y, et les profils-colonnes de celle de x :

$$f_{k|h} = \frac{n_{hk}}{n_{h\bullet}} \approx f_{\bullet k} = \frac{n_{\bullet k}}{n} \text{ et } \quad f_{h|k} = \frac{n_{hk}}{n_{\bullet k}} \approx f_{h\bullet} = \frac{n_{h\bullet}}{n}$$

Contribution au χ^2

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

C'est équivalent à dire que les fréquences conjointes doivent être proches des produits de fréquences marginales.

$$f_{hk} = \frac{n_{hk}}{n} \approx f_{h\bullet}\, f_{\bullet k} = \frac{n_{h\bullet}}{n}\,\frac{n_{\bullet k}}{n}$$

Contribution au χ^2

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

Un des moyens de quantifier leur proximité est de calculer la distance du χ^2 de l'une par rapport à l'autre. Dans ce cas particulier, on parle de χ^2 de contingence .

Contribution au χ^2

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

La distance du chi-deux de contingence de la distribution empirique (f_{hk}) à la distribution théorique $(f_{h\bullet}.f_{\bullet k})$ vaut :

$$D_{\chi^2} = \sum_{h=1}^{r} \sum_{k=1}^{s} \frac{(f_{hk} - f_{h\bullet} f_{\bullet k})^2}{f_{h\bullet} f_{\bullet k}}$$

Contribution au χ^2

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

La distance du chi-deux de contingence de la distribution empirique (f_{hk}) à la distribution théorique $(f_{h\bullet}.f_{\bullet k})$ vaut :

$$D_{\chi^2} = -1 + \sum_{h=1}^{r}\sum_{k=1}^{s}\frac{(n_{nk}^2)}{n_{h\bullet}n_{\bullet k}}$$

Contribution au χ^2

$X \setminus Y$	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
...	...		...		...	...
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
...	...		...		...	...
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

Le lien évolue de 0 (indépendance totale) à 1 (dépendance totale).

Analyse Factorielle des Correspondances (AFC)

Introduction

L'analyse factorielle peut être vu comme une ACP (Analyse en Composantes Principales) mais qui utilise une métrique χ^2 pondérée.

Pourquoi parle-t-on de "Correspondances" :

- Variable numérique : Corrélation
- Variable nominale : Correspondances

Pourquoi "*Factorielle*" :

- Décomposition du tableau de contingence en une somme de tableaux qui sont le produit de facteurs simples.

Introduction

L'analyse factorielle des correspondances vise à rassembler en un nombre réduit de dimensions la plus grande partie de l'information initiale en s'attachant non pas aux valeurs absolues mais aux correspondances entre les variables, c'est-à-dire aux valeurs relatives.

L'AFC offre la particularité (contrairement aux ACP) de fournir un espace de représentation commun aux variables et aux individus.

Introduction

Méthodologie

- Matrice des fréquences (r lignes, s colonnes) et $r <= s.f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}$
- Matrice de liaison. $M_{ij} = \frac{f_{ij} - f_{i\bullet}.f_{\bullet j}}{\sqrt{f_{i\bullet}.f_{\bullet j}}}$
- V = M'M (matrice basse) ou MM' (matrice haute) et X' est la transposée de la matrice X.
- Soit D, la diagonalisation de la matrice V. Recherche des valeur propres de V. $Det(V - \lambda I) = 0$
- Projection de M par D. Produit matricielle : AFC = M'D

Tableau de contingence

	P1	P2	P3	P4	P5
R1	28	14	45	33	12
R2	36	21	25	64	23
R3	21	64	38	11	7
R4	79	42	67	9	41

Matrice des fréquences (M)

	P1	P2	P3	P4	P5
R1	0.041176471	0.020588235	0.066176471	0.048529412	0.017647059
R2	0.052941176	0.030882353	0.036764706	0.094117647	0.033823529
R3	0.030882353	0.094117647	0.055882353	0.016176471	0.010294118
R4	0.116176471	0.061764706	0.098529412	0.013235294	0.060294118

V = M'M

	R1	R2	R3	R4
R1	0.261941351	0.203307819	0.252463142	0.182346888
R2	0.203307819	0.287629263	0.23523261	0.140405753
R3	0.252463142	0.23523261	0.275094933	0.183209496
R4	0.182346888	0.140405753	0.183209496	0.287907509

Décomposition en valeurs propres

Résolution de cette équation : $Det(V - \lambda I) = 0$

	R1	R2	R3	R4
R1	$\lambda 1$	0	0	0
R2	0	$\lambda 2$	0	0
R3	0	0	$\lambda 3$	0
R4	0	0	0	$\lambda 4$

Mais $\lambda 1$ est toujours égal à 1.
Dans notre cas :

- $\lambda 2 = 0.155$ (61.425 % d'info)
- $\lambda 3 = 0.080$(31.7 % d'info)
- $\lambda 4 = 0.017$(6.799 % d'info)

La somme des valeurs propres est égale au χ^2

Attention au choix des axes principaux.

Décomposition en valeurs propres (matrice D)

Résolution de cette équation : $Det(V - \lambda I) = 0$

	R1	R2	R3	R4
R1	$\lambda 1$	0	0	0
R2	0	$\lambda 2$	0	0
R3	0	0	$\lambda 3$	0
R4	0	0	0	0

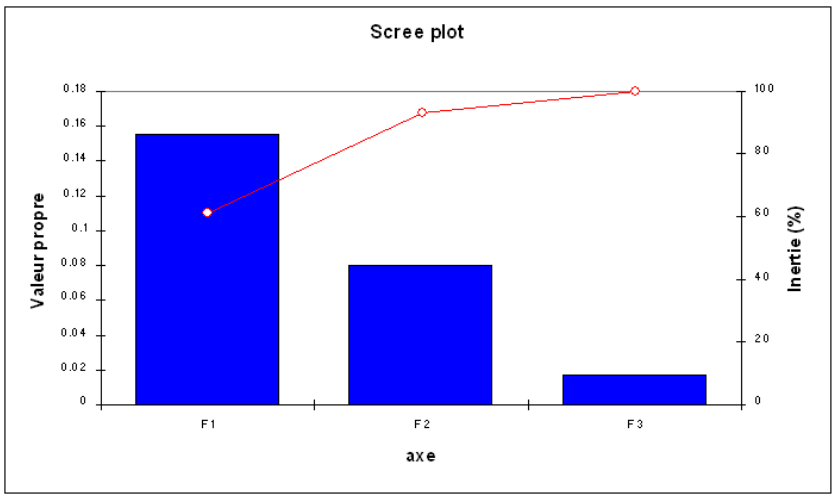


Figure: Inertie des valeurs propres

Projection des produits(variables) sur les axes factoriels

Produit matriciel : $M'.D$

- $P1_{axe1} = (M'_{11}.\lambda 1) + (M'_{21} * 0) + ... + (M'_{41} * 0))$
- $P1_{axe2} = (M'_{11} * 0) + (M'_{21} * \lambda 2) + ... + (M'_{41} * 0))$
- ....
- $P1_{axe4} = (M'_{11} * 0) + (M'_{21} * 0) + ... + (M'_{41} * \lambda 4))$

La première analyse est alors terminée. Nous l'avons dit plus haut nous nous trouvons dans un cas de relations duales, il est inutile de faire la seconde analyse, les coordonnées des 4 points "régions" se déduisent immédiatement de celles des cinqs points "produits"

Projection des régions(individus) sur les axes factoriels

Cette projection se déduit de la manière suivante: Pour l'individu 1 : (R1)

$$i1_{axe1} = (1/K_{i1} * \sqrt{\lambda 1}) * (K_{1,1} * V1_{axe1} + K_{1,2} * V2_{axe1} \ldots K_{1,n} * Vn_{axe1})$$
$$i1_{axe1} = (1/K_{i1} * \sqrt{\lambda 2}) * (K_{1,1} * V1_{axe2} + K_{1,2} * V2_{axe2} \ldots K_{1,n} * Vn_{axe2})$$
....
$$ip_{axek} = (1/K_{ip} * \sqrt{\lambda k}) * (K_{p,1} * V1_{axek} + K_{p,2} * V2_{axek} \ldots K_{p,n} * Vn_{axek})$$

- où : i1, i2 ... ip sont les p individus de la série brute de départ.
- axe1, axe2 ... axek sont les k axes d'inertie ;
- $\lambda 1$, $\lambda 2$... λk sont k valeurs propres ;
- $K_{i,j}$ sont les données brutes de la séries ;
- Vn_{axek} sont les coordonnées des n variables sur les k axes d'inertie.

Projection des individus et de s produits sur les axes factoriels

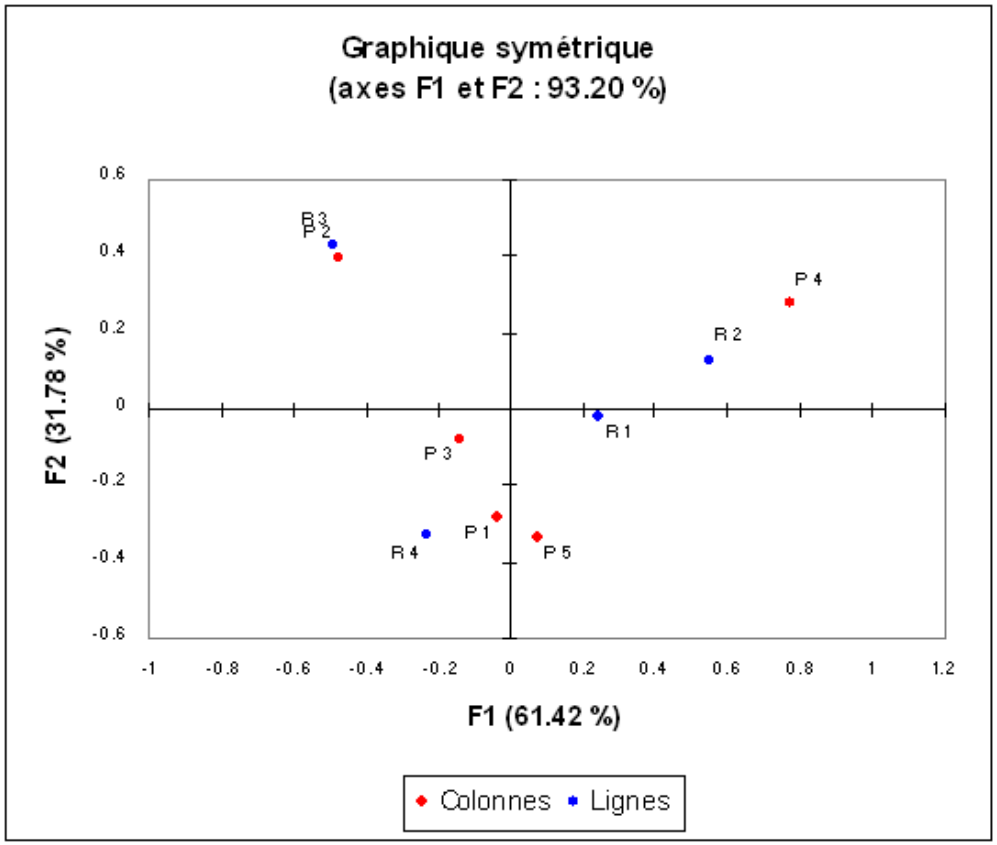


Figure: Projection symétrique des individus et des variables sur les 2 premiers axes factoriels

Projection des individus et de s produits sur les axes factoriels

La proximité entre un point-ligne L et un point-colonne C ne possède pas d'interprétation géométrique immédiate. En revanche, l'angle de sommet O (centre du plan factoriel) et dont les côtés passent par L et C a la propriété suivante :

- si l'angle (OL, OC) est aigu, la modalité-ligne L et la modalité colonne C s' "attirent " : produit scalaire positif (taux de liaison positif)
- si l'angle (OL, OC) est obtus, la modalité-ligne L et la modalité colonne C se "repoussent" (taux de liaison négatif) : Produit scalire négatif.
- si l'angle (OL, OC) est droit, la modalité-ligne L et la modalité colonne C n'interagissent pas (taux de liaison voisin de 0).

Conclusion

- AFC pour l'analyse de données nominales.
  - Etude de tableau de contingence.
  - Visualisation dans un espace 2D décorrélé de tableaux de grandes dimensions.
  - Description des interactions entre les variables.
- Les inconvénients ?
  - N'appréhende que les relations affines entre les variables. (phénoméne linèaire)

Références (liens)

- http://rb.ec-lille.fr/Cours_de_recueil_analyse_et_traitement_de_donnees.htm
- http://web.univ-pau.fr/RECHERCHE/SET/LAFFLY/docs_laffly/INTRODUCTION_AFC.pdf
- http://www.cnam-econometrie.com/upload/afclecon.pdf
- http://fr.wikipedia.org/wiki/Analyse_factorielle_des_correspondances